ELSEVIER

Ecological Modelling 154 (2002) 135–150

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks

Julian D. Olden *, Donald A. Jackson

Department of Zoology, University of Toronto, 25 Harbord Street, Toronto, Ontario, Canada M5S 3G5

Received 2 May 2000; received in revised form 15 February 2002; accepted 4 March 2002

Abstract

With the growth of statistical modeling in the ecological sciences, researchers are using more complex methods, such as artificial neural networks (ANNs), to address problems associated with pattern recognition and prediction. Although in many studies ANNs have been shown to exhibit superior predictive power compared to traditional approaches, they have also been labeled a "black box" because they provide little explanatory insight into the relative influence of the independent variables in the prediction process. This lack of explanatory power is a major concern to ecologists since the interpretation of statistical models is desirable for gaining knowledge of the causal relationships driving ecological phenomena. In this study, we describe a number of methods for understanding the mechanics of ANNs (e.g. Neural Interpretation Diagram, Garson's algorithm, sensitivity analysis). Next, we propose and demonstrate a randomization approach for statistically assessing the importance of axon connection weights and the contribution of input variables in the neural network. This approach provides researchers with the ability to eliminate null-connections between neurons whose weights do not significantly influence the network output (i.e. predicted response variable), thus facilitating the interpretation of individual and interacting contributions of the input variables in the network. Furthermore, the randomization approach can identify variables that significantly contribute to network predictions, thereby providing a variable selection method for ANNs. We show that by extending randomization approaches to ANNs, the "black box" mechanics of ANNs can be greatly illuminated. Thus, by coupling this new explanatory power of neural networks with its strong predictive abilities, ANNs promise to be a valuable quantitative tool to evaluate, understand, and predict ecological phenomena.

Keywords: Connection weights; Sensitivity analysis; Neural Interpretation Diagram; Garson's algorithm; Statistical models

* Corresponding author. Present address: Department of Biology, Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO 80523-1878, USA. Tel.: +1-970-491-5432; fax: +1-970-491-0649.

E-mail addresses: olden@lamar.colostate.edu (J.D. Olden), jackson@zoo.utoronto.ca (D.A. Jackson).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00064-9

1. Introduction

Artificial neural networks (ANNs) are receiving increased attention in the ecological sciences as a powerful, flexible, statistical modeling technique for uncovering patterns in data (Colasanti, 1991; Edwards and Morse, 1995; Lek et al., 1996a; Lek and Guégan, 2000); a fact recently demonstrated during the first international workshop on the applications of neural networks to ecological modeling (conference papers are published in *Ecological Modelling*: Volume 120, Issue 2–3). The utility of ANNs for solving complex pattern recognition problems has been demonstrated in many terrestrial (e.g. Paruelo and Tomasel, 1997; Özesmi and Özesmi, 1999; Manel et al., 1999; Spitz and Lek, 1999) and aquatic studies (e.g. Lek et al., 1996a,b; Bastarache et al., 1997; Mastrorillo et al., 1998; Chen and Ware, 1999; Gozlan et al., 1999; Olden and Jackson, 2001; Scardi, 2001), and has led many researchers to advocate ANNs as an attractive, non-linear alternative to traditional statistical methods.

The primary application of ANNs involves the development of predictive models to forecast future values of a particular response variable from a given set of independent variables. Although the predictive value of ANNs appeals greatly to many ecologists, researchers have often criticized the explanatory value of ANNs, calling it a "black box" approach to modeling ecological phenomena (e.g. Paruelo and Tomasel, 1997; Lek and Guégan, 1999; Özesmi and Özesmi, 1999). This view stems from the fact that the contribution of the input variables in predicting the value of the output is difficult to disentangle within the network. Consequently, input variables are often entered into the network and an output value is generated without gaining any understanding of the inter-relationships between the variables, and therefore, providing no explanatory insight into the underlying mechanisms being modeled by the network (Anderson, 1995; Bishop, 1995; Ripley, 1996). The "black box" nature of ANNs is a major weakness compared to traditional statistical approaches that can readily quantify the influence of the independent variables in the modeling process, as well as provide a measure of the degree of confidence regarding their contribution. Currently, there is a lack of theoretical or practical ways to partition the contributions of the independent variables in ANNs (Smith, 1994); thus presenting a substantial drawback in the ecological sciences where the interpretation of statistical models is desirable for gaining insight into causal relationships driving ecological phenomena.

Recently, a number of methods have been proposed for selecting the best network architecture (i.e. number of neurons and topology of connections) among a set of candidate networks, e.g. asymptotic comparison techniques, approximate Bayesian analysis, and cross validation (Dimopoulos et al., 1995; see Bishop, 1995 for review). In contrast, methods for quantifying the independent variable contributions within networks are more complicated, and as a result are rarely used in ecological studies. For example, intensive computational approaches such as growing and pruning algorithms (Bishop, 1995), partial derivatives (e.g. Dimopoulos et al., 1995, 1999) and asymptotic t-tests are often not used in favour of simpler techniques that use network connection weights (e.g. Garson's algorithm: Garson, 1991; Lek's algorithm: Lek et al., 1996a). Although these simpler approaches provide a means of determining the overall influence of each predictor variable, interactions among the variables are more difficult to interpret since the strength and direction of individual axon connection weights within the network must be examined directly. Bishop (1995) discusses the use of pruning algorithms to remove connection weights that do not contribute to the predictive performance of the neural network. In brief, a pruning approach begins with a highly connected network (i.e. large number of connections among neurons), and then successively removes weak connections (i.e. small absolute weights) or connections that cause a minimal change in the network error function when removed. A logical question then arises: at what threshold value (i.e. absolute connection weight or change in network error) should weights be removed or retained in the network? In the present study, we propose a randomization test for ANNs to address this question. This randomization approach provides a statistical pruning technique for

eliminating null connection weights that minimally influence the predicted output, as well as provides a selection method for identifying independent variables that significantly contribute to network predictions. By using randomization protocols to partition the importance of connection weights (in terms of their magnitude and direction), researchers will be able to quantitatively assess both the individual and interactive effects of the input variables in the network prediction process, as well as evaluate the overall contributions of the variables. Using an empirical example describing the relationship between fish species richness and habitat characteristics of north-temperate lakes, we illustrate the utility of the ANN randomization test, and compare its results to two commonly used approaches: Garson's algorithm and sensitivity analysis.

2. Empirical example: fish species richness–habitat relationships in lakes

Throughout this paper we use an empirical example relating fish species richness to habitat conditions of 286 freshwater lakes located in Algonquin Provincial Park, south-central Ontario, Canada (45°50′ N, 78°20′ W). We tabulated species presence for each lake to examine relationships between fish species richness (ranging from 1 to 23) and a suite of habitat-related variables (8 in total). Predictor variables were chosen to include factors that have been shown to be related to critical habitat requirements of fish in this geographic region (Minns, 1989), including: surface area, lake volume, and total shoreline perimeter which are correlated with habitat diversity; maximum depth which is negatively correlated with winter dissolved-oxygen concentrations and related to thermal stratification; surface measurements (taken at depths ≤ 2.0 m) of pH and total dissolved solids to provide an estimate of nutrient status and lake productivity; lake elevation which is related to both habitat heterogeneity and colonization/extinction features of the lake; and growing degree-days which is a surrogate for productivity.

3. Interpreting neural-network connection weights: an important caveat

We refrain from detailing the specifics of neural network optimization and design, and instead refer the reader to the extensive coverage provided in the texts by Smith (1994), Bishop (1995), Ripley (1996), as well as articles by Ripley (1994), Cheng and Titterington (1994). It is sufficient to say that the methods described in this paper refer to the classic family of one hidden-layer, feed-forward neural network trained by the backpropagation algorithm (Rumelhart et al., 1986). These neural networks are commonly used in ecological studies because they are suggested to be universal approximators of any continuous function (Hornik et al., 1989). *N*-fold cross validation was used to determine optimal network design as it provides a nearly unbiased estimate of prediction success (Olden and Jackson, 2000). We found that a neural network with four hidden neurons exhibited good predictive power ($r = 0.72$ between observed and predicted species richness).

In the neural network, the connection weights between neurons are the links between the inputs and the outputs, and therefore are the links between the problem and the solution. The relative contributions of the independent variables to the predictive output of the neural network depend primarily on the magnitude and direction of the connection weights. Input variables with larger connection weights represent greater intensities of signal transfer, and therefore are more important in the prediction process compared to variables with smaller weights. Negative connection weights represent inhibitory effects on neurons (reducing the intensity of the incoming signal) and decrease the value of the predicted response, whereas positive connection weights represent excitatory effects on neurons (increasing the intensity of the incoming signal) and increase the value of the predicted response.

Given the obvious importance of connection weights in assessing the relative contributions of the independent variables, there is one topic that we believe warrants additional attention. During the optimization process, it is necessary that the network converges to the global minimum of the

fitting criterion (e.g. prediction error) rather than one of the many local minima. Connection weights in networks that have converged to a local minimum will differ from networks that have globally converged, thus resulting in the misinterpretation of variable contributions. Two approaches can be employed to ensure the greatest probability of network convergence to the global minimum. The first approach involves combining different local minima rather than choosing between them, for example, by averaging the outputs of networks using the connection weights corresponding to different local minima (e.g. Wolpert, 1992; Perrone and Cooper, 1993; Ripley, 1995). This approach would also presumably involve averaging the contributions of input variables across the networks representing each of the local minima. The second approach employs global optimization procedures where parameters such as learning rate, momentum or regularization are used during network training (e.g. White, 1989; Gelfand and Mitter, 1991; Ripley, 1994). The addition of a learning rate (η) and momentum (α) parameters during optimization has been used in the ecological literature (e.g. Lek et al., 1996a; Mastrorillo et al., 1997a; Gozlan et al., 1999; Spitz and Lek, 1999; Olden and Jackson, 2001) because in addition to reducing the problem of convergence to local minima, it also accelerates the optimization process. The η regulates the magnitude of changes in the weights and biases during optimization, and α mediates the contribution of the last weight change in the previous iteration to the weight change in the current iteration. The values of η and α can be set constant or can vary during network optimization, although there are a number of the disadvantages to holding η and α constant (see Bishop, 1995). Consequently, values of both η and α are commonly modified by either increasing or decreasing their value according to whether the error decreased or increased, respectively, during the previous iteration of network optimization (e.g. Hagan et al., 1996; Mastrorillo et al., 1998; Özesmi and Özesmi, 1999). In our study, we included learning rate and momentum parameters during the optimization process (defining them as a function of the error), and started the network optimization with random connections weights between -0.3 and 0.3. The variable learning rate and momentum parameters, and the small interval of initial random weights ensured a high probability of global network convergence and thus provided greater confidence regarding the validity of the connection weights and their interpretation.

4. Illuminating the "black box"

4.1. Preparation of the data

Prior to neural network optimization, the data set must be transformed so that the dependent and independent variables exhibit particular distributional characteristics. The dependent variable must be converted to the range [0…1] so that it conforms to the demands of the transfer function used (sigmoid function) in the building of the neural network. This is accomplished by using the formula:

$$r_n = \frac{y_n - \min(Y)}{\max(Y) - \min(Y)} \quad (1)$$

where r_n is the converted response value for observation n, y_n is the original response value for observation n, and $\min(Y)$ and $\max(Y)$ represent the minimum and maximum values, respectively, of the response variable Y. Note that the dependent variable does not have to be converted when modeling a binary response variable (e.g. species presence/absence) because its values already fall within this range.

To standardize the measurement scales of the network inputs, the independent variables are converted to z-scores (i.e. mean = 0, standard deviation = 1) using the formula:

$$z_n = \frac{x_n - \overline{X}}{\sigma_X} \quad (2)$$

where z_n is the standardized value of observation n, x_n is the original value of observation n, and $\overline{X}$ and σ_X are the mean and standard deviation of the variable X. It is essential to standardize the input variables so that same percentage change in the weighted sum of the inputs causes a similar percentage change in the unit output. Both the dependent and independent variables of the rich-

ness–habitat data set were modified using the above formulas.

4.2. Methods for quantifying input variable contributions in ANNs

In the following section, we detail a series of methods that are available to aid in the interpretation of connection weights and variable contributions in neural networks. These approaches have been used by ecologists and represent a set of techniques for understanding neuron connections in networks. Next, we extend a randomization approach to these methods, illustrating how connection weights and the overall influence of the input variables in the network can be assessed statistically.

4.2.1. Neural Interpretation Diagram (NID)

Recently, a number of investigators have advocated using axon connection weights to interpret predictor variable contributions in neural networks (e.g. Aoki and Komatsu, 1999; Chen and Ware, 1999). Özesmi and Özesmi (1999) proposed the Neural Interpretation Diagram (NID) for providing a visual interpretation of the connection weights among neurons, where the relative magnitude of each connection weight is represented by line thickness (i.e. thicker lines representing greater weights) and line shading represents the direction of the weight (i.e. black lines representing positive, excitator signals and gray lines representing negative, inhibitor signals). Tracking the magnitude and direction of weights between neurons enables researchers to identify individual and interacting effects of the input variables on the output. Fig. 1 illustrates the NID for the empirical example and shows the relative influence of each habitat factor in predicting fish species richness. The relationship between the inputs and outputs is determined in two steps since there are input-hidden layer connections and hidden-output layer connections. Positive effects of input variables are depicted by positive input-hidden and positive hidden-output connection weights, or negative input-hidden and negative hidden-output connection weights. Negative effects of input variables are depicted by positive input-hidden and negative hidden-output connection

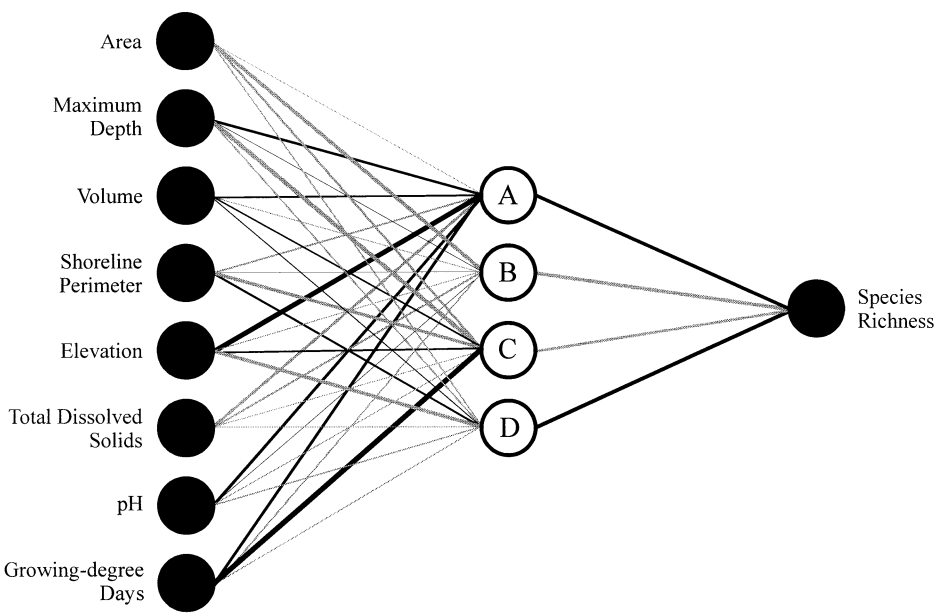


Fig. 1. NID for neural network modeling fish species richness as a function of eight habitat variables. The thickness of the lines joining neurons is proportional to the magnitude of the connection weight, and the shade of the line indicates the direction of the interaction between neurons: black connections are positive (excitator) and gray connections are negative (inhibitor).

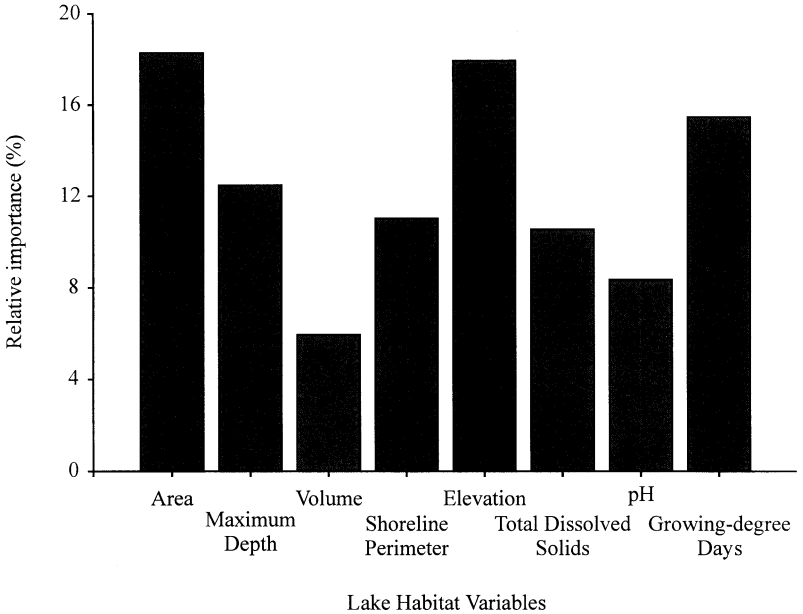


Fig. 2. Bar plots showing the percentage relative importance of each habitat variable for predicting fish species richness based on Garson's algorithm (Garson, 1991). See Box 1 for sample calculations.

weights, or by negative input-hidden and positive hidden-output connection weights. Therefore, the multiplication of the two connection weight directions (positive or negative) indicates the effect that each input variable has on the output variable. Interactions among predictor variables can be identified as input variables with opposing connection weights entering the same hidden neuron.

The interpretation of connection weights, and more specifically NIDs, is not an easy task because of the complexity of connections among the neurons (Fig. 1). Additional hidden neurons would only make this interpretation more difficult. Furthermore, a subjective choice must be made regarding the magnitude at which connection weight should be interpreted. These considerations make the direct examination of connection weights challenging at best and virtually impossible in data sets with large numbers of variables. We show later that a randomization approach can aid in the interpretation NIDs by identifying non-significant connection weights that can be removed.

4.2.2. Garson's algorithm

Garson (1991) proposed a method, later modified by Goh (1995), for partitioning the neural network connection weights in order to determine the relative importance of each input variable in the network (see Box 1 for a summary of the protocol). This approach has been used in a number of ecological studies, including Mastrorillo et al. (1997b, 1998), Gozlan et al. (1999), Aurelle et al. (1999), Brosse et al. (1999, 2001). It is important to note that Garson's algorithm uses the absolute values of the connection weights when calculating variable contributions, and therefore does not provide the direction of the relationship between the input and output variables. Fig. 2 illustrates the results from our empirical example, highlighting the relative importance of the habitat variables. Predictor contributions ranged from 6 to 18%, with lake area and elevation exhibiting the strongest relationships with predicted species richness, and lake volume and pH showing the weakness relationship.

4.2.3. Sensitivity analysis

A number of investigators have used sensitivity analysis to determine the spectrum of input variable contributions in neural networks. Recently, a number of alternative types of sensitivity analysis have been proposed in the ecological literature. For example, the Senso-nets approach includes an additional weight in the network for each input variable representing the variable's sensitivity (Schleiter et al., 1999). Scardi and Harding (1999) added white noise to each input variable and examined the resulting changes in the mean square error of the output. Traditional sensitivity analysis involves varying each input variable across its entire range while holding all other input variables constant; so that the individual contributions of each variable are assessed. This approach is somewhat cumbersome, however, because there may be an overwhelming number of variable combinations to examine. As a result, it is common first to calculate a series of summary measures for each of the input variables (e.g. minimum, maximum, quartiles, percentiles), and then vary each input variable from its minimum to maximum value, in turn, while all other variables are held constant at each of these measures (e.g. Özesmi and Özesmi, 1999). Relationships between each input variable and the response can be examined for each summary measure, or the calculated response can be averaged across the summary measures. Holding the input variables constant at a small number of values provides a more manageable sensitivity analysis, yet still requires a great deal of the time because each value of the input variable must be examined. Consequently, Lek et al. (1995, 1996a,b) suggested examining only 12 data values delimiting 11 equal intervals over the variable range rather than examining its entire range (this has been termed Lek's algorithm). Contribution plots can be constructed by averaging the response value across all summary statistics for each of the 12 values of the input variable of interest. Many studies have employed Lek's algorithm, including Lek et al. (1995, 1996a), Mastrorillo et al. (1997a, 1998), Guégan et al. (1998), Laë et al. (1999), Lek-Ang et al. (1999), Spitz and Lek (1999). In this study, we constructed contribution plots for each of the eight predictor variables in the neural network by varying each input variable across its entire range and holding all other variables constant at their 20th, 40th, 60th and 80th percentile (Fig. 3). It is evident from the contribution plots that the influence of the habitat variables on predicted species richness in the study lakes varies greatly depending on what summary value the other input variables are held. Although the predicted output may exhibit any number of relationships with the independent variables, below is a summary of the response curves observed in our empirical example.

- Gaussian response curve—input variable contributes greatest at intermediate values, and exhibits decreasing influence at low and high values: e.g. influence of pH and growing-degree days on species richness.
- Bimodal response curve—input variable contributes greatest at low and high values, and exhibits minimal influence at intermediate values: e.g. influence of surface area, maximum depth, shoreline perimeter and total dissolved solids on species richness when all other variables are low in value.
- Left-skewed response curve—input variable contributes greatest at high values, and exhibits minimal influence at low and intermediate values: e.g. influence of lake elevation on species richness.
- Right-skewed response curve—input variable contributes greatest at low values, and exhibits minimal influence at intermediate and high values: e.g. influence of total dissolved solids on species richness, influence of overall lake size (i.e. surface area, maximum depth, volume and shoreline perimeter) on species richness when all other variables are intermediate in value.
- Decreasing response curve—input variable contributes decreasingly at increasing values: e.g. influence of surface area on species richness when all other variables are high in value.
- Flat response curve—input variables contributes minimally across its entire range: e.g. influence of growing-degree days on species richness when all other variables are high in value.

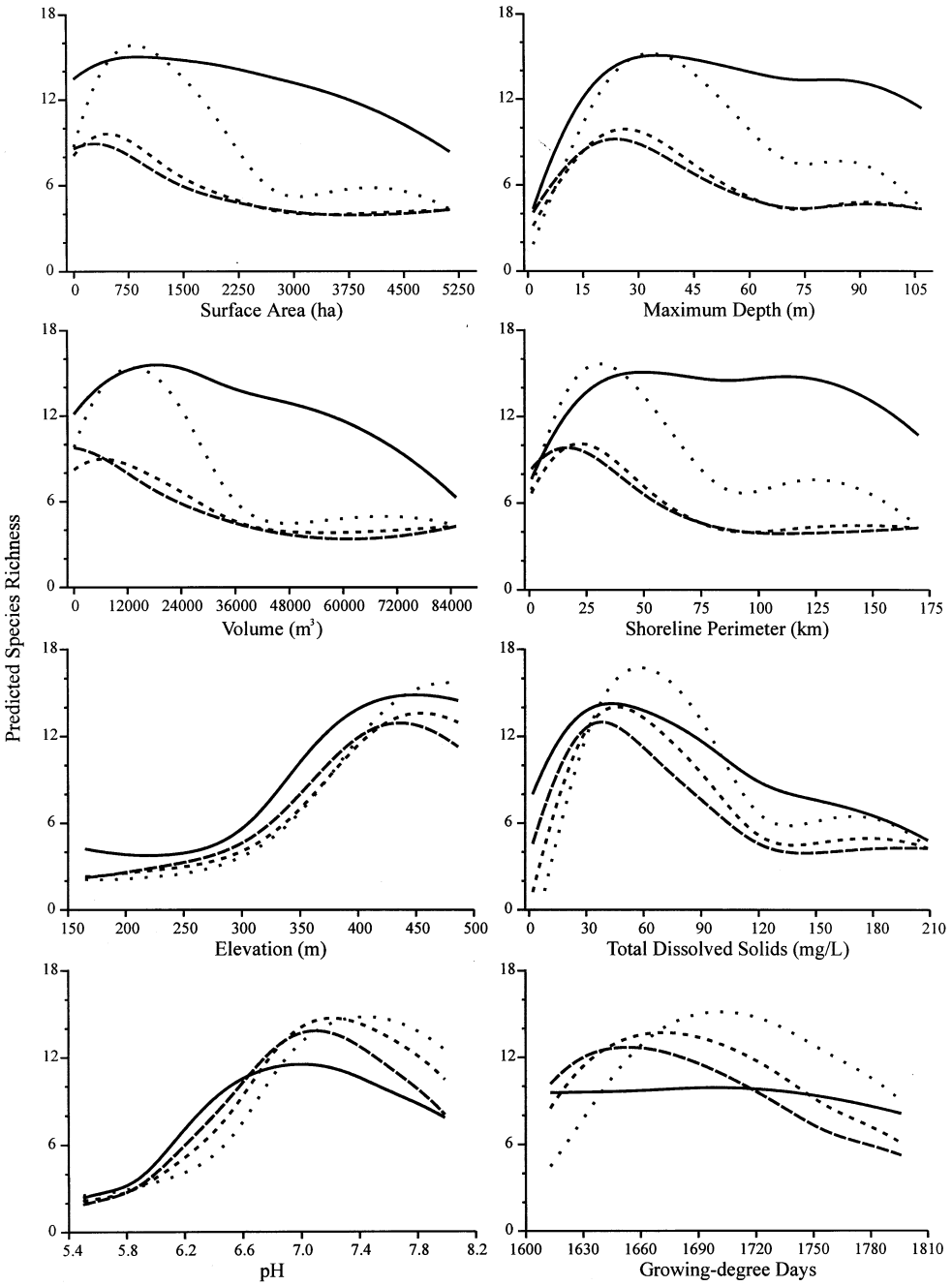


Fig. 3. Contribution plots from the sensitivity analysis illustrating the neural network response curves to changes in each habitat variable with all other variables held at their 20th (· · · ·), 40th (-----), 60th (- - -) and 80th (—) percentile.

4.2.4. Randomization test for artificial neural networks

We propose a randomization test for input-hidden-output connection weight selection in neural networks. By eliminating null-connection weights that do not differ significantly from random, we can simplify the interpretation of neural networks by reducing the number of axon pathways that have to be examined for direct and indirect (i.e. interaction) effects on the response variable, for instance when using NIDs. This objective is similar to statistical pruning techniques (e.g. asymptotic t-tests), yet does not have to conform to the assumptions of parametric and non-parametric methods because the randomization approach empirically constructs the distribution of expected values under the null hypothesis for the test statistic (i.e. weight connection) from the data at hand. Moreover, the randomization approach can be used as a variable selection method for ANNs by summing across input-hidden-output connection weights or calculating the relative importance (i.e. Garson's algorithm) for each input variable. This approach provides a quantitative tool for selecting statistically significant input variables for inclusion into the network, again reducing network complexity and assisting in the network interpretation. The following is the randomization protocol for testing the statistical significance of connection weights and input variables:

1. construct a number of neural networks using the original data with different initial random weights;
2. select the neural network with the best predictive performance, record initial random connection weights used in constructing this network, and calculate and record:
   (a) input-hidden-output connection weights: the product of input-hidden and hidden-output connection weights for each input and hidden neuron (e.g. observed c_{A1}: step 2, Box 1);
   (b) overall connection weight: the sum of the input-hidden-output connection weights for each input variable (e.g. observed $c_1 = c_{A1} + c_{B1}$);
   (c) relative importance (%) for each input variable based on Garson's algorithm (e.g. observed RI_1: step 4, Box 1);
3. randomly permute the original response variable (y_{random});
4. construct a neural network using y_{random} and the initial random connection weights; and
5. repeat steps (3) and (4) a large number of times (i.e. 999 times in this study) each time recording 2(a), (b) and (c); e.g. randomized c_{A1}, randomized c_1, and randomized RI_1.

The statistical significance of each input-hidden-output connection weight, overall connection weight and relative importance of each input variable (e.g. observed c_{A1}, observed c_1 and observed RI_1) can be calculated as the proportion of randomized values (e.g. randomized c_{A1}, randomized c_1 and randomized RI_1), including the observed, whose value is equal to or more extreme than the observed values. Fig. 4 illustrates the distribution of randomized input-hidden-output connection weights (for hidden neuron B), overall connection weight and relative importance of surface area for predicting species richness of lakes.

Table 1 contains the connection weight structure for the neural network and the associated p-values from the randomization tests. The results show that only a fraction of the total 32 input-hidden-output connections (i.e. 8 inputs $\times$ 4 hidden neurons) are statistically different from what would be expected based on chance alone. For instance, only six input-hidden-output connections are significant at $\alpha = 0.05$. The results also show that when you account for all connection weights (i.e. overall connection weight), lake size (i.e. surface area, maximum depth, volume and shoreline perimeter) and pH are positively associated with species richness, while elevation, total dissolved solids and growing-degree days are negatively associated with species richness. However, only the influence of maximum depth and shoreline perimeter are statistically significant (Table 1). Interestingly, the results from the randomization test using relative importance (derived from Garson's algorithm) differ from the results using overall connection weights. Using Garson's algorithm, surface area was the only significant factor correlated with species richness, and elevation was marginally nonsignificant (Table 1). The discrepancy between the two approaches results from the different ways that the methods use the

Table 1
Axon connection weights for the neural network modeling fish species richness as a function of eight habitat variables

i	Predictor variable	Hidden neuron A		Hidden neuron B		Hidden neuron C		Hidden neuron D		Overall connection weight		Relative importance	
		W_{Ai}	*P*	W_{Bi}	*P*	W_{Ci}	*P*	W_{Di}	*P*	$\Sigma W_{Ai...+...Di}$	*P*	%	*P*
1	Area (ha)	−0.92	0.275	*7.81*	*0.003*	3.25	0.183	−2.71	0.094	7.43	0.087	*18.27*	*0.031*
2	Max. depth (m)	3.76	0.063	−0.60	0.475	6.23	0.051	−2.58	0.123	*6.81*	*0.002*	12.49	0.511
3	Volume (m^3)	2.58	0.110	0.92	0.236	−1.76	0.156	0.80	0.359	2.54	0.301	5.94	0.885
4	Sh. per. (km)	−2.64	0.102	0.28	0.432	4.95	0.053	3.45	0.088	*6.04*	*0.021*	11.02	0.585
5	Elevation (m)	*8.10*	*0.024*	0.44	0.458	−2.55	0.322	*−7.16*	*0.012*	−1.17	0.217	17.94	0.084
6	TDS (mg/l)	*−4.54*	*0.043*	3.88	0.138	0.71	0.334	−0.49	0.519	−0.44	0.433	10.54	0.577
7	pH	*4.15*	*0.049*	−0.67	0.268	1.31	0.331	−2.39	0.177	2.40	0.067	8.34	0.808
8	GDD	4.36	0.098	2.49	0.117	*−7.43*	*0.005*	−1.45	0.361	−2.03	0.114	15.46	0.194

W_{ij} represents the input-hidden-output connection weight for hidden neuron *i* (where *i* = A–D) and input variable *j* (where *j* = 1–8). *P* values for input-hidden-output connection weights (W_{Ai}, W_{Bi}, W_{Ci}, and W_{Di}), overall connection weights ($\Sigma W_{Ai...+...Di}$), and Garson's relative importance (%) are based on 999 randomizations. Italics indicate statistical significance based on $\alpha = 0.05$. Abbreviations Sh. per., TDS and GDD refer to shoreline perimeter, total dissolved solids, and growing-degree days, respectively.

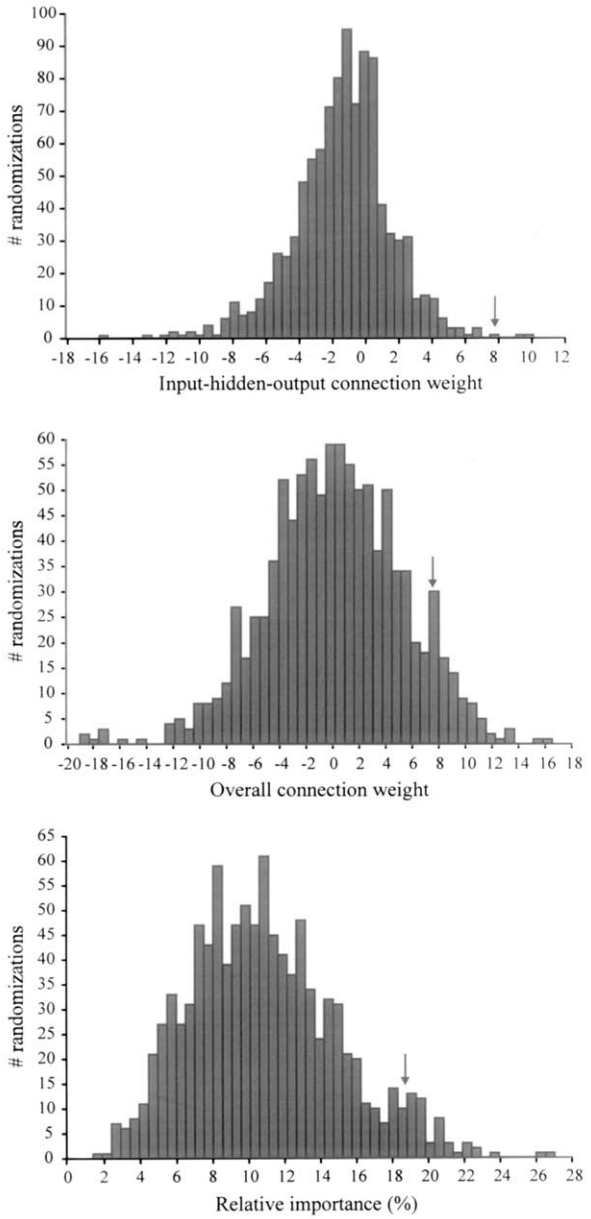


Fig. 4. Distributions of random input-hidden-output connection weights for hidden neuron B, overall connection weight, and input relative importance (%) for the influence of surface area on lake species richness. Arrows represent observed input-hidden-output connection weight for hidden neuron B (7.81), overall connection weight (7.43) and relative importance (18.27%).

network connection weights. Garson's algorithm uses absolute connection weights to calculate the influence of each input variable on the response (see Box 1), whereas overall connection weight is calculated using the original values. Examining Fig. 5 ($\alpha = 0.05$), we can show that Garson's algorithm can be potentially misleading for the interpretation of input variable contributions. It is evident that lake elevation shows a strong, positive association with species richness through hidden neuron A, but a strong, negative relationship with richness through hidden neuron D. Based on absolute weights, Garson's algorithm indicates a large relative importance of that variable because both connection weights have large magnitudes (e.g. for all hidden neurons RI = 17.94%: Table 1). However, in such a case the influence of the input variable on the response is actually negligible since the positive influence through hidden neuron A is counteracted by the negative influence through hidden neuron D (e.g. for all hidden neurons $\Sigma W_{A1\ldots D1} = -1.17$, $P = 0.217$: Table 1). For this reason, we believe caution should be employed when making inferences from the results generated by Garson's algorithm since the direction of the input–output interaction is not taken into account.

Using results of the randomization test, we removed non-significant connection weights from the NID (originally shown in Fig. 1), resulting in Fig. 5 which illustrates only connection weights that were statistically significantly different from random at $\alpha = 0.05$ and $\alpha = 0.10$. Focusing on hidden neuron C in Fig. 5 ($\alpha = 0.10$), it is apparent that as maximum depth and shoreline perimeter increase, and growing-degree days decreases, species richness increases in the study lakes. Furthermore, interactions among habitat factors can be identified as input variables with contrasting connection weights (i.e. opposite directions) entering the same hidden neuron. For example, in examining hidden neuron D it is evident that lake shoreline perimeter interacts with lake elevation. An increase in lake elevation decreases predicted species richness; however, this negative effect weakens as shoreline perimeter increases. Therefore, there is an interaction between lake elevation and shoreline perimeter in that high elevation lakes with convoluted shorelines have greater spe-

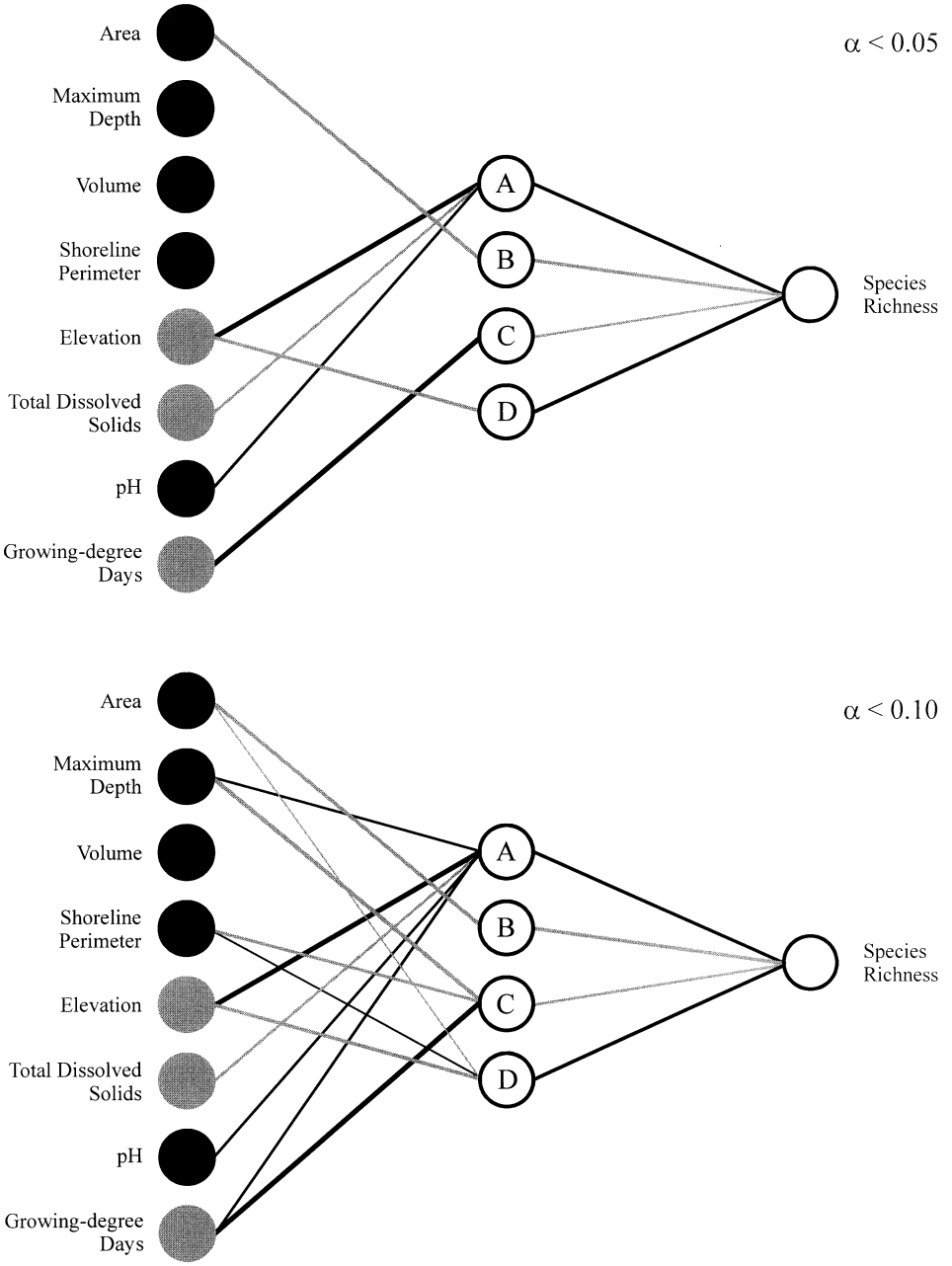
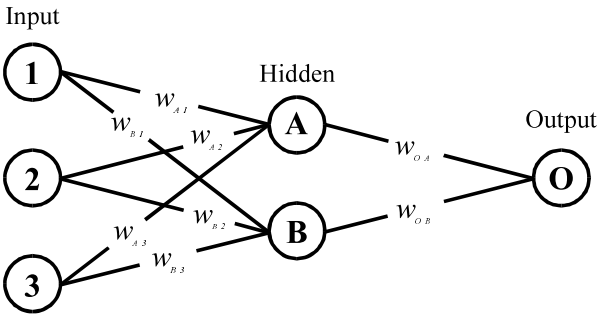


Fig. 5. NID after non-significant input-hidden-output connection weights are eliminated using the randomization test (i.e. connection weights statistically different from zero based on $\alpha = 0.05$ and $\alpha = 0.10$). The thickness of the lines joining neurons is proportional to the magnitude of the connection weight, and the shade of the line indicates the direction of the interaction between neurons: black connections are positive (excitator) and gray connections are negative (inhibitor). Black input neurons indicate habitat variables that have an overall positive influence on species richness, and gray input neurons indicate an overall negative influence on species richness (based on overall connection weights).



1. Matrix containing input-hidden-output neuron connection weights

	Hidden A	Hidden B
Input 1	w_{A1} = -2.61	w_{B1} = -1.23
Input 2	w_{A2} = 0.13	w_{B2} = -0.91
Input 3	w_{A3} = -0.69	w_{B3} = -2.09
Output	w_{OA} = 1.11	w_{OB} = 0.39

2. Contribution of each input neuron to the output via each hidden neuron calculated as the product of the input-hidden connection and the hidden-output connection: e.g., $c_{A1} = w_{A1}$ x w_{OA} = -2.61 x 1.11 = -2.90

	Hidden A	Hidden B
Input 1	c_{A1} = -2.90	c_{B1} = -0.48
Input 2	c_{A2} = 0.14	c_{B2} = -0.35
Input 3	c_{A3} = -0.77	c_{B3} = -0.82

3. Relative contribution of each input neuron to the outgoing signal of each hidden neuron: e.g., $r_{A1} = |c_{A1}|/(|c_{A1}|+|c_{A2}|+|c_{A3}|)$= 2.90 / (2.90 + 0.14 + 0.77) = 0.76; and sum of input neuron contributions: e.g., $S_1 = r_{A1} + r_{B1}$ = 0.76 + 0.29 = 1.05

	Hidden A	Hidden B	Sum
Input 1	r_{A1} = 0.76	r_{B1} = 0.29	S_1 = 1.05
Input 2	r_{A2} = 0.04	r_{B2} = 0.21	S_2 = 0.25
Input 3	r_{A3} = 0.20	r_{B3} = 0.50	S_3 = 0.70

4. Relative importance of each input variable: e.g., $RI_1 = S_1 / (S_1 + S_2 + S_3)$ x 100 = 1.05 / (1.05 + 0.25 + 0.70) x 100 = 52.5 %

	Relative importance
Input 1	52.5 %
Input 2	12.5 %
Input 3	35.0 %

Box 1. Garson's algorithm for partitioning and quantifying neural network connection weights. Sample calculations shown for three input neurons (1, 2 and 3), two hidden neurons (A and B), and one output neuron (○).

cies richness compared to high elevation lakes with simple shorelines. The NID also identifies input variables that do not interact, for example lake volume, because this variable does not exhibit significant weights with contrasting effects at any single hidden neuron with any of the other variables.

The randomization test can also be used as a variable selection method for removing input and hidden neurons for which incoming or outgoing connection weights are not significantly different from random. For example, no significant weights originate from maximum depth, volume and shoreline perimeter input neurons in Fig. 5 (α = 0.05), indicating that these variables do not contribute significantly to predicted values of species richness. These neurons (i.e. predictor variables) could be removed from the analysis with little loss

of predictive power. Similarly, hidden neurons lacking connections to significant weights could also be removed (this does not occur in our empirical example, but see Olden and Jackson, 2001). In summary, a randomization approach to neural networks can aid greatly in the identification and interpretation of direct and indirect (i.e. interaction between input variables) contributions of input variables in ANNs. We refer the reader to Olden (2000), Olden and Jackson (2001) for additional studies using the randomization test.

Two important components of the randomization test involved the optimization of the neural network. First, we conducted the randomization test for the product of input-hidden and hidden-output weights rather than each input-hidden and hidden-output connection weight separately, because the direction of the connection weights (i.e. positive or negative) can switch between different networks optimized with the same data (i.e. symmetric interchanges of weights: Ripley, 1994). For instance, the input-hidden and hidden-output weights might both be positive in one network and both negative in another, but in both cases the input variable exerts a positive influence on the response variable. To remove this problem, we examined the product of the input-hidden-output weights because the true direction of the relationship between the input and output will be conserved. Second, optimizing the neural network several times (i.e. constructing a number of networks with the same data but different initial random weights) can result in neural networks with identical predictive performance, but quite different connection weights. Therefore, if different initial random weights are used for each randomization, dissimilarities between the observed and random connection weights cannot be differentiated from the differences arising solely from different initial connection weights. Using the same initial random weights for each randomized network alleviates this problem.

5. Conclusion

We reiterate the concern raised by a number of ecologists and explicit to our paper: are ANNs a black box approach for modeling ecological phenomena? In light of the synthesis provided here, we argue the answer is unequivocally no. We have reviewed a series of methods, ranging from qualitative (i.e. NIDs) to quantitative (i.e. Garson's algorithm and sensitivity analysis), for interpreting neural-network connection weights, and have demonstrated the utility of these methods for shedding light on the inner workings of neural networks. These methods provide a means for partitioning and interpreting the contribution of input variables in the neural network modeling process. In addition, we described a randomization procedure for testing the statistical significance of these contributions in terms of individual connection weights and overall influence of each input variable. The former case facilitates the interpretation of direct and interacting effects of input variables on the response by removing connection weights that do not contribute significantly to the performance of the neural network. In the latter case, the randomization test assesses whether the contribution of a particular input variable on the response differs from what would be expected by chance. The randomization procedure enables the removal of null neural pathways and non-significant input variables; thereby aiding in the interpretation of the neural network by reducing its complexity. In conclusion, by coupling the explanatory insight of neural networks with its powerful predictive abilities, ANNs have great promise in ecology, as a tool to evaluate, understand, and predict ecological phenomena.

Acknowledgements

We thank Sovan Lek for his insightful comments regarding the finer points of artificial neural networks, and for providing some of the original MatLab code for this study. This manuscript was greatly improved by the comments of Nick Collins, Brian Shuter, Keith Somers and an anonymous reviewer. Funding for this research was provided by a Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) to J.D. Olden, and an NSERC Research Grant to D.A. Jackson.

References

Anderson, J.A., 1995. An Introduction to Neural Networks. MIT, Cambridge, MA, 650 pp.

Aoki, I., Komatsu, T., 1999. Analysis and prediction of the fluctuation of sardine abundance using a neural network. Oceanol. Acta 20, 81–88.

Aurelle, D., Lek, S., Giraudel, J.-L., Berrebi, P., 1999. Microsatellites and artificial neural networks: tools for the discrimination between natural and hatchery brown trout (*Salmo trutta*, L.) in Atlantic populations. Ecol. Model. 120, 313–324.

Bastarache, D., El-Jabi, N., Turkkan, N., Clair, T.A., 1997. Predicting conductivity and acidity for small streams using neural networks. Can. J. Civ. Eng. 24, 1030–1039.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Brosse, S., Guégan, J.F., Tourenq, J.N., Lek, S., 1999. The use of neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecol. Model. 120, 299–311.

Brosse, S., Lek, S., Townsend, C.R., 2001. Abundance, diversity, and structure of freshwater invertebrates and fish communities: an artificial neural network approach. New Zealand J. Mar. Fresh. Res. 35, 135–145.

Chen, D.G., Ware, D.M., 1999. A neural network model for forecasting fish stock recruitment. Can. J. Fish. Aquat. Sci. 56, 2385–2396.

Cheng, B., Titterington, D.M., 1994. Neural networks: a review from a statistical perspective (with discussion). Stat. Sci. 9, 2–54.

Colasanti, R.L., 1991. Discussions of the possible use of neural network algorithms in ecological modeling. Binary 3, 13–15.

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization. Neural Process. Lett. 2, 1–4.

Dimopoulos, I., Chronopoulos, J., Chronopoulos-Sereli, A., Lek, S., 1999. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). Ecol. Model. 120, 157–165.

Edwards, M., Morse, D.R., 1995. The potential for computer-aided identification in biodiversity research. Trend Ecol. Evol. 10, 153–158.

Garson, G.D., 1991. Interpreting neural-network connection weights. Artif. Intell. Expert 6, 47–51.

Gelfand, S.B., Mitter, S.K., 1991. Recursive stochastic algorithms for global optimization in Rd. SIAM J. Control Optimization 29, 999–1018.

Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. 9, 143–151.

Gozlan, R.E., Mastrorillo, S., Copp, G.H., Lek, S., 1999. Predicting the structure and diversity of young-of-the-year fish assemblages in large rivers. Fresh. Biol. 41, 809–820.

Guégan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384.

Hagan, M.T., Demuth, H.B., Beale, M.H., 1996. Neural Network Design. PWS Publishing, Boston, MA.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Laë, R., Lek, S., Moreau, J., 1999. Predicting fish yield of African lakes using neural networks. Ecol. Model. 120, 325–335.

Lek, S., Beland, A., Dimopoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Mar. Freshwater Res. 46, 1229–1236.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996b. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Resour. 9, 23–29.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996a. Application of neural networks to modeling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modeling, an introduction. Ecol. Model. 120, 65–73.

Lek, S., Guégan, J.F. (Ed.), 2000. Artificial Neuronal Networks, Applications to Ecology and Evolution. Springer-Verlag, New York, USA.

Lek-Ang, S., Deharveng, L., Lek, S., 1999. Predictive models of collembolan diversity and abundance in a riparian habitat. Ecol. Model. 120, 247–260.

Manel, S., Dias, J.-M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120, 337–347.

Mastrorillo, S., Lek, S., Dauba, F., 1997a. Predicting the abundance of minnow *Phoxinus phoxinus* (Cyprinidae) in the River Ariege (France) using artificial neural networks. Aquat. Living Resour. 10, 169–176.

Mastrorillo, S., Lek, S., Dauba, F., Beland, A., 1997b. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Fresh. Biol. 38, 237–246.

Mastrorillo, S., Dauba, F., Oberdorff, T., Guégan, J.F., Lek, S., 1998. Predicting local fish species richness in the Garonne River basin. C.R. Acad. Sci. Paris, Sciences de la vie 321, 423–428.

Minns, C.K., 1989. Factors affecting fish species richness in Ontario lakes. Trans. Am. Fish. Soc. 118, 533–545.

Olden, J.D., 2000. An artificial neural network approach for studying phytoplankton succession. Hydrobiology 436, 131–143.

Olden, J.D., Jackson, D.A., 2000. Torturing data for the sake of generality: how valid are our regression models? Écoscience 7, 501–510.

Olden, J.D., Jackson, D.A., 2001. Fish-habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. Trans. Am. Fish. Soc. 130, 878–897.

Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modeling with interspecific interaction. Ecol. Model. 116, 15–31.

Paruelo, J.M., Tomasel, F., 1997. Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecol. Model. 98, 173–186.

Perrone, M.P., Cooper, L.N., 1993. When networks disagree: Ensemble methods for hybrid neural networks. In: Mammone, R.J. (Ed.), Artificial Neural Networks for Speech and Vision. Chapman and Hall, London, pp. 126–147.

Ripley, B.D., 1994. Neural networks and related methods for classification. J. R. Stat. Soc. B 56, 409–456.

Ripley, B.D., 1995. Statistical ideas for selecting network architectures. In: Kappen, B., Gielen, S. (Eds.), Neural Networks: Artificial Intelligence and Industrial Applications. Springer, London, pp. 183–190.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.

Scardi, M., 2001. Advances in neural network modeling of phytoplankton primary production. Ecol. Model. 146, 33–46.

Scardi, M., Harding, L.W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol. Model. 120, 213–223.

Schleiter, I.M., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K.-D., Schmidt, H.-H., Werner, H., 1999. Modeling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. Ecol. Model. 120, 271–286.

Smith, M., 1994. Neural Networks for Statistical Modeling. Van Nostrand Reinhold, New York, USA.

Spitz, F., Lek, S., 1999. Environmental impact prediction using neural network modeling. An example in wildlife damage. J. Appl. Ecol. 36, 317–326.

White, H., 1989. Learning in artificial neural networks: a statistical perspective. Neural Computing 1, 425–464.

Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241–259.